Income Statistics Division

62F0026MIE - 01003

Methodology of the Survey of Household Spending

Prepared by:
Sophie Arsenault
Johanne Tremblay

October 2001

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Client Services, Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; income@statcan.ca) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialing area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no.62F0026MIE-01003, is available on internet for free. Users can obtain single issues at: http://www.statcan.ca/cgi-bin/downpub/research.cgi.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Income Statistics Division

Methodology of the Survey of Household Spending

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2001

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

October 2001

Catalogue no. 62F0026MIE - 01003

Frequency: Irregular

Ottawa

La version française de cette publication est disponible sur demande

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Abstract

This document provides a detailed description of the methodology of the Survey of Household Spending. Topics covered include: target population; sample design; data collection; data processing; weighting and estimation; estimation of sampling error; and data suppression and confidentiality.

ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

TABLE OF CONTENTS

1. INTRODUCTION 9

2. TARGET POPULATION 10

3. SAMPLE DESIGN 10
3.1 Size and allocation of the SHS sample 10
3.2 LFS sample design (cluster selection) 12
3.3 Selection of the SHS sample 14
3.4 Specifics of the design for the territories 15

4. DATA COLLECTION 15
4.1 Data collection methodology 16
4.2 Interviews and follow-up procedures 17
4.3 Supervision and controls 17
4.4 Non-response to the SHS 18

5. DATA PROCESSING 18
5.1 Coding and data entry 18
5.2 Edit and imputation 18

6. WEIGHTING AND ESTIMATION 20
6.1 Sampling weight 20
6.2 Non-response adjustment 20
6.3 Adjustment using auxiliary information 21
6.4 Adjustment for influential data 22
6.5 Estimates 23

7. ESTIMATION OF SAMPLING ERROR 23
7.1 Model for approximating the CV for domain estimates 24
7.2 Model for approximating the CV based on the microdata file 24

8. DATA SUPPRESSION AND CONFIDENTIALITY 24
8.1 Suppression of unreliable data in estimate tables 24
8.2 Confidentiality of microdata files 25

9. CHANGES IN THE SURVEY METHODOLOGY 25

BIBLIOGRAPHY 26

APPENDIX 1
Formula for computation of the variance of estimates by the jackknife method 27

APPENDIX 2
Formula for approximating the CV for a domain (a population subgroup) 28

ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

106-01985/

1. INTRODUCTION

The Survey of Household Spending (SHS) is an annual survey that collects information from Canadian households about spending habits, dwelling characteristics and household equipment.

The SHS was first conducted in January 1998 to collect household spending data for 1997. It replaced the periodic Family Expenditure Survey (FAMEX), generally conducted every four years.[1] The annual survey was introduced to meet a need for more accurate and more frequent provincial data for the National Accounts. As a result, it uses a larger sample. Another major change brought in with the new survey was a complete redesign of the questionnaire. The amount of detail required on expenditures was reduced substantially, which decreased interview times by about 33%.

The SHS has many objectives, as it serves as a valuable data source for a number of Statistics Canada products and for many users outside the Agency. For example, SHS estimates are used in preparing National Accounts estimates of personal spending on goods and services at the national and provincial levels. Data provided by the survey are also used to update the basket used in computing the Consumer Price Index. They are used to set low-income cut-offs, which are in turn used by the Survey of Labour and Income Dynamics to determine the percentages of low-income individuals and families. In addition, SHS data are one of the five data sources used to develop the Social Policy Simulation Model, which analyzes the impact of various economic and social policies. Since the Household Facilities and Equipment Survey was discontinued, the SHS has become the source of data on dwelling characteristics and household equipment needed by Canada Mortgage and Housing Corporation and many outside users.

SHS data are collected through personal interviews conducted with a sample of households in Canada's ten provinces and three territories. The households are contacted early in the year, between January and March, and are asked about their expenditures for the previous calendar year. The data are then edited and weighted. Output from the survey includes tables and microdata files needed by the various users.

This document provides a detailed description of the survey methodology: sample design, data collection and processing, production of estimates and other products, and dissemination rules. For a more general description of the methodology, consult the Users' Guide for each of the survey years [1].

[1] The last Family Expenditure Survey covered the 1996 reference year. National data were also collected in 1992, 1986, 1982 and 1978. In some years, such as 1990 and 1984, the survey was conducted in major cities only.

2. TARGET POPULATION

The target population of the SHS consists of individuals living in Canadian private households, excluding official representatives of foreign countries living in Canada and their families, and residents of Indian reserves and crown lands. The "private households" restriction means that the target population also excludes residents of institutions, such as prisons, chronic-care hospitals and senior citizens' residences; members of religious orders and other communal colonies; members of the Armed Forces living in military camps; and individuals living permanently in hotels and rooming houses.

The survey covers nearly 98% of the population of the ten provinces. In the Yukon, Northwest Territories and Nunavut, the coverage is 81%, 92% and 89% of the population, respectively (or 80%, 93% and 90% of households). Note that in these regions, individuals living in very small communities (generally consisting of fewer than 100 households) or in unorganized areas are excluded from the target population.

Until 1996, coverage of the territories by the periodic Family Expenditure Survey was limited to the cities of Whitehorse and Yellowknife. In the 1997 Survey of Household Spending, that coverage was extended to 78% for Yukon and 70% for the Northwest Territories and Nunavut. The current coverage was achieved with the 1998 survey. Starting with the 2000 SHS, the territories are surveyed only once every two years to reduce the response burden.

3. SAMPLE DESIGN

Expenditure data are collected from a sample of households selected according to a multi-stage, stratified sampling design. This design varies according to the level of urbanization, but generally consists of a two-stage sample for which the first level is an area sample, i.e., a sample of geographic areas called clusters. In the second stage, dwellings are selected from a list of all private dwellings in the selected clusters. All the selected dwellings that are inhabited by individuals from the target population constitute the survey sample.

To minimize operating costs, the SHS uses largely the same sample design as the Labour Force Survey (LFS). The dwellings in the SHS sample are selected from LFS sample clusters, but the two surveys use different dwelling samples. The main aspects of the LFS cluster sample design are described in section 3.2. A more detailed description is available in the LFS methodology publication [2]. The characteristics used in selecting SHS dwellings from LFS sample clusters are covered in section 3.3. The sample design used in the territories is different; its specifics are described in section 3.4. The section immediately below explains how the sample is allocated among the provinces and territories.

3.1 Size and allocation of the SHS sample

The size of the SHS sample may vary from year to year. When the survey was launched in 1997, the sample size was set at about 24,000 households, some

67% more than in the FAMEX sample. Since then, the sample size has fluctuated up or down slightly depending on budgetary pressures and whether the territories were included in the survey. The total sample size for each year is shown in Table 3.1.

Each year, the total sample is allocated among the provinces and territories (when the latter are included) so as to obtain estimates of similar reliability. More specifically, the sample allocation is based on the variability of income in each province: a larger proportion of the sample is allocated to provinces where the difference between the highest and lowest incomes is greater. Population size is also taken into account, though to a much lesser extent. For the territories and Prince Edward Island, which has a much smaller population than the other provinces, the sample size is predetermined to ensure that the sample does not contain an excessive proportion of the population. At present, the SHS samples about 4% of the population in each territory and 2% of the population in Prince Edward Island.

Response rates for previous surveys are used to adjust the sample size of each province and territory. Vacancy rates (the proportion of unoccupied dwellings) are used for the same purpose at the provincial and subprovincial levels. The rates are provided by the most recent LFS data for the January-to-March period corresponding to the SHS collection period.

Lastly, each provincial or territorial sample is distributed in direct proportion to the size of the population in the census metropolitan areas.[2] The proportion of the sample allocated outside the census metropolitan areas matches the LFS allocation [2], except in the Northwest Territories and Nunavut, where the LFS is not conducted.

Table 3.1 presents the SHS's provincial and territorial sample sizes for each year since 1997 in terms of the number of households (excluding selected dwellings that were unoccupied or out of scope). In the 1997 survey, a larger proportion of the sample was allocated to Newfoundland, Nova Scotia and New Brunswick because we expected the sample size to increase in subsequent years and we wanted to assign the increased sample right away to the provinces that had signed the goods and services tax harmonization agreement. It turned out later that the survey budget limited the sample size to a maximum of 24,000 households, and the sample allocation was adjusted in subsequent years.

[2] And for the cities of Charlottetown, Summerside, Whitehorse, Yellowknife and Iqaluit.

Table 3.1
Sample size (number of households) by province or territory

Provinces and Territories	Sample size (number of households)			
	SHS 1997	SHS 1998	SHS 1999	SHS 2000[3]
Canada	**23,842**	**20,236**	**23,518**	**20,877**
Newfoundland	1,997	1,343	1,937	1,794
P.E.I.	795	807	822	822
N.S.	2,424	1,573	2,199	2,040
N.B.	2,044	1,406	1,957	1,821
Quebec	3,122	2,848	2,710	2,516
Ontario	3,362	3,056	3,453	3,202
Manitoba	1,772	1,739	2,034	1,882
Saskatchewan	1,478	1,721	1,837	1,697
Alberta	2,743	2,186	2,519	2,336
B.C.	3,010	2,590	2,985	2,768
Total, provinces	**22,747**	**19,269**	**22,453**	**20,877**
Yukon	451	383	403	0
N.W.T.	644	383	414	0
Nunavut		201	248	0
Total, territories	**1,095**	**967**	**1,065**	**0**

3.2 LFS sample design (cluster selection)

The LFS sample design is based on data from the Census of Population and is redesigned after each decennial census to reflect changes in the population. The current design is based on 1991 Census data.

The principles underlying the LFS sample design are the same for every province. First, each province is divided into a number of geographic regions based on the intersections of economic regions (ERs) and Employment Insurance Economic Regions (EIERs). In particular, every census metropolitan area forms a geographic region since it is an EIER.

Each geographic region is then divided into types of areas, primarily urban areas, rural areas and remote areas. The sample design varies according to the type of area.

Urban areas

In some major cities with large numbers of apartment buildings, both an apartment list frame and an area frame are used. In other urban areas, only an area frame is employed.

[3] For 2000, the figures are approximations based on the vacancy rate and the in-scope rate in the previous survey. The exact number of households in the sample is not known until interviewers have visited the selected dwellings and eliminated those which are unoccupied or are occupied by out-of-scope individuals.

An area frame is a list of geographic zones making up each area. These zones are combined to form strata. There can be up to three levels of stratification. At the top levels, the aim is generally to form geographically compact and contiguous strata, whereas at the bottom level, the requirement is for final strata that are as homogeneous as possible with respect to certain socio-economic characteristics. In a few large cities,[4] separate strata are formed from enumeration areas with high average household incomes (about $100,000 or more).

To reduce collection costs, the households that make up the final stratum are not selected directly. Instead, the stratum is divided into clusters. In urban areas, the clusters may be combinations of block-faces, enumeration areas (EAs) or parts of EAs. Then clusters are selected (usually six, sometimes a multiple of six) in each stratum with a probability proportional to cluster size. For example, if one cluster is twice as large as another is, the former will be twice as likely to be selected as the latter.

The apartment list frame is a list of apartments prepared using information from the Canada Mortgage and Housing Corporation. This frame provides better representation of apartment residents and minimizes the effect of cluster growth due to construction of new apartment buildings. In some cities,[5] apartment strata are divided into two categories: low-income (where the average household income is under $20,000) and regular. For each stratum in the frame, apartment buildings are selected for the first-stage sample with a probability proportional to the number of apartments in the building.

In low-density urban areas, which are highly dispersed towns, a different sample design is used. Sampling is done in three stages: first, towns are selected within the strata; then, clusters (block-faces) are chosen within the towns; and finally, dwellings are selected within the clusters.

Rural areas

Only an area frame is used in rural areas. Geographic strata are formed by combining two or three census divisions, which are then subdivided, where numbers permit, to form strata that are homogeneous with respect to certain socio-economic characteristics. In the first stage of sampling, enumeration areas are selected within each final stratum with a probability proportional to the number of households in the EA.

In low-density rural areas, a variation on the sample design is used. Two or three primary sampling units consisting of a group of six EAs are selected in the first stage, and then a sample of dwellings is selected within each of the EAs of the selected primary sampling units.

[4] Montreal, Ottawa, Toronto, Hamilton, London, Winnipeg, Calgary and Vancouver.
[5] Montreal, Ottawa-Hull, Toronto, Winnipeg, Calgary, Edmonton and Vancouver.

Remote areas

The northern parts of the provinces (excluding the Maritimes) are, for the most part, sparsely populated. Samples for those areas are usually selected in two steps. First, a sample of EAs and of agglomerations known as places is selected. Three-stage sampling is also used in one remote area in Quebec.

Places with fewer than 10 households or 25 persons are excluded from the sample design, as are EAs with fewer than 25 households. Despite these exclusions, the design covers 90% of the population of remote areas in the provinces.

3.3 Selection of the SHS sample

Interviewers visit the clusters selected in the LFS sample design and make a list of all the private dwellings they contain. From that list, one sample is chosen for the LFS and a different one is selected for the SHS. Dwellings are selected by systematic sampling.

Since the SHS uses a much smaller sample than the LFS, dwellings are not selected in every LFS cluster. The LFS is a panel survey in which households remain in the sample for six months. The LFS sample was designed so that it could be divided into six representative subsamples to permit rotation of one sixth of the sample each month. That is why six clusters (or a multiple of six) are selected in each final stratum, one per rotation group. This method makes it easy to select a smaller sample for another survey since a subset of the rotation groups can be used. This is generally the approach used for LFS supplementary surveys. For the SHS, the number of rotation groups is determined at the stratum level according to the survey's specific needs with regard to provincial and subprovincial allocation of the sample. In some instances, only part of a rotation group is needed. Where that is the case, households are randomly removed from the sample.

The dwelling sample is obtained following the cluster listing operation. Since the sampling rates are predetermined, there may be a difference between the expected and actual sample sizes if the number of dwellings on the list differs from the number used in developing the survey's sample design. To keep collection costs in check (since cluster sizes tend to increase) and prevent significant disparities in interviewer workloads, two methods are used to control the sample size.

The problem is usually corrected by randomly removing some of the originally selected dwellings. This process of keeping the sample size at the desired level is known as sample stabilization. When the number of dwellings increases sharply in certain urban areas, cluster subsampling is used instead. There are three options, depending on how large the increase is and how similar the new dwellings are to others in the same stratum: form subclusters; create a new stratum; or subsample the dwellings in the cluster.

3.4 Specifics of the design for the territories

Since the SHS has to cover all of the territories, unlike FAMEX, which included only Whitehorse and Yellowknife, a new sample design was introduced for the territories in the 1998 survey.[6] This design was different because it had to reflect the fact that a large portion of the population is scattered among low-density communities. This characteristic has a major impact on the survey's collection costs. Hence, individuals living in unorganized areas, very small communities (generally fewer than 100 households) or inaccessible areas, are excluded from the survey's target population.

Despite this difference in coverage levels, which results in the exclusion from the SHS of about 19% of the Yukon's population, the method used in this territory is similar to the one used in the provinces since the LFS is also conducted in the Yukon.

In the Northwest Territories and Nunavut, a specific design was developed for the SHS since the LFS did not collect information there. That design is based on 1996 Census data, whereas the LFS design used for Yukon and the provinces is based on 1991 Census data.

The cities of Yellowknife and Iqaluit each form a separate stratum divided into clusters, with a sample of dwellings selected in each cluster. Other communities are combined into two or three strata on the basis of socio-demographic characteristics such as population size, proportion of Native People and average household income. Each community makes up a cluster, and a sample of two or three clusters is selected in each stratum. Then a sample of about 30 dwellings is chosen in each of the selected clusters.

SHS data for the 2000 reference year were not collected in the three territories, following the decision to survey them only every other year so as to ease the heavy response burden on their populations.

4. DATA COLLECTION

The SHS collects information about the entire budget of Canadian households on a voluntary basis. This information includes expenditures, income, and changes in assets and debts over the 12-month period from January 1 to December 31 of the reference year.

The SHS also gathers information about dwelling characteristics and the household equipment owned by households. This information reflects the situation on December 31 of the reference year.

[6] The 1997 survey was transitional, as the probability selection method for the territories had not yet been finalized. On the basis of an arbitrary choice of communities and the sample for Whitehorse and Yellowknife, we were able to produce estimates representing 78% of Yukon's population and 70% of the Northwest Territories and Nunavut combined.

4.1 Data collection methodology

Interviewers collect the data in face-to-face interviews with respondents. The interviews are conducted in the first three months of the year following the survey's reference year.

The SHS is a recall survey, which means that respondents have to remember the expenditures they made during the one-year reference period. To reduce the recall effort and help them provide more accurate information, respondents are encouraged to consult records relating to the reference period, such as mortgage statements, cheque registers, credit card account statements, and income tax returns. For items purchased at regular intervals, information is generally collected by asking respondents the quantity they bought, the frequency of the purchases, and the typical price, and then those figures are used to derive estimates of annual expenditures for the household. In particular, annual figures for food expenditures are usually derived from data collected over a short period (a week or a month). The SHS collects only total food expenditures. Detailed data are collected once every four years by means of an expenditure diary in the Food Expenditure Survey.

The SHS questionnaire collects information about the household, such as expenditures for housing, furniture, food, transportation and recreation. Some information must be supplied for individual household members, such as personal income, taxes and clothing expenditures. This information is often obtained by proxy.

Members of a household

To obtain expenditure data for a household, we must first accurately identify its members. The person or group of people who occupies a dwelling constitutes a household. For the SHS, the members of that household are defined as follows:

i) all persons living in the dwelling at the time of the interview who have no permanent residence elsewhere and are not members of another household;
ii) any persons who were members of the household during the reference period, or part of the reference period, even if they do not live there at the time of the interview.

In reporting a household's income and expenditure, it is important to include the income and expenditure of members who have left the household and those who joined the household during the reference year.[7] In the latter case, the data must reflect the portion of the year during which the individual was a member of the household.

[7] For persons who joined a household, it is necessary to determine whether they were previously living in a household that no longer exists. If so, the former household had no chance of being selected. The data are collected on a different questionnaire for the portion of the reference period preceding the change in households.

Another possibility is that a household existed for only part of a reference year. That is the case, for example, where two young adults living with their parents get married and form a new household during the reference period. The expenditures of such households cover only part of the reference year. Such households receive special treatment in the computation of certain estimates. This point is discussed at greater length in section 5.5.

4.2 Interviews and follow-up procedures

The interviews are conducted by Statistics Canada interviewers, many of whom also collect information for the LFS. The interviewers receive special training for the SHS.

A week before visiting, the interviewer sends the occupants of the selected dwellings a letter of introduction emphasizing the survey's importance and the confidentiality of the information collected. He or she then visits the household to conduct the interview. If the timing is inconvenient, the interviewer makes an appointment to return at a more convenient time. If there is no one home, many additional attempts are made to contact the household, for example, visiting at different times of day, or consulting reverse directories to find out the occupant's telephone number.

Because a wide range of information is needed, lengthy interviews may be required in some cases, and the interviewer may have to visit more than once to get all the information. On average, the interview takes about one hour and forty minutes. At the end of the interview, respondents may keep a summary of the expenditures they have reported for their own records.

If a person refuses to take part in the SHS, the Regional Office mails a letter to the dwelling stressing the importance of the survey and the household's cooperation. Next, the interviewer makes a second visit (or call). If the interviewer is unable to persuade the household to take part, he or she will prepare a non-interview report. Depending on the comments provided, the senior interviewer will decide whether to make further refusal conversion attempts.

4.3 Supervision and controls

All SHS interviewers report to senior interviewers, who are responsible for ensuring that the interviewers are familiar with the survey's concepts and methodology, for periodically monitoring their work, and for reviewing completed documents. Senior interviewers are supervised by program managers working at Statistics Canada's regional offices.

The interviewer carries out the initial edit, making sure that the information is complete in all sections of the questionnaire. The questionnaires are then checked by senior interviewers.

Since respondents' recall is a key component of the quality of SHS data, one of the controls involves measuring the difference between receipts (income and other money received by the household) and disbursements (expenditures plus

net change in assets and liabilities) reported by the household. If the difference is more than 10% of the larger of receipts or disbursements, the interviewer or senior interviewer will contact the respondents again to obtain further information and attempt to identify errors or omissions.

4.4 Non-response to the SHS

Despite all the effort put into collecting the information, there are always some non-respondent households. For example, contact could not be made; unusual circumstances such as illness or death prevented the interview; or the household members refused to take part in the survey. For each survey year, the report on data quality contains detailed information about the non-response rates [3]. The collection non-response rates for the past few years are shown in Table 4.1. Also shown in the table are the final non-response rates, which take into account those households, which were excluded following data processing (see section 5.4).

Table 4.1
SHS non-response rate

Reference year	Collection non-response rate			Final non-response rate (at estimation)
	TOTAL	No contact	Refusal	
1997	20.7	5.8	15.0	24.4
1998	20.7	4.9	15.8	23.6
1999	23.6	5.9	17.7	26.8

5. DATA PROCESSING

The main steps in the processing of SHS data are response coding, data entry, editing, imputation of partial non-response, identification of usable data, and weighting. The latter will be covered in section 6.

5.1 Coding and data entry

Very few questions in the SHS require coding. Coding is done by the interviewer and checked by the senior interviewer. Then the questionnaires are put into batches of 20 and the data is keyed in at Statistics Canada's regional offices. Data entry is checked by selecting a sample of questionnaires from each data entry operator for rekeying. If the number of errors for a questionnaire exceeds a certain threshold, the entire batch is sent back for rekeying. The size of the sample selected for editing depends on the past performance of the data entry operators.

5.2 Edit and imputation

The first step in the automated edit process is carried out after each questionnaire has been checked manually by the interviewer and the senior interviewer. There are a number of "must pass" rules that check for consistency

between answers in the questionnaire. The edit also identifies unusual situations that might require correction. This part of the automated edit is done at Statistics Canada's regional offices, so that respondents can be contacted if additional information is needed to resolve inconsistencies in their responses. Problems identified during this edit are dealt with by members of specially trained questionnaire resolution teams. Subsequently, the data are transmitted to Head Office for further editing and correction of invalid responses.

In cases of partial non-response (where the respondent has failed to answer only some of the questions), the missing data are imputed. The imputation method depends on whether the data are categorical or continuous. Categorical variables can take only specific values (such as yes-or-no questions and type-of-dwelling questions), whereas continuous variables can take any numerical value (such as income and expenditure).

Categorical data, which occur mostly in the dwelling characteristics and household equipment sections of the questionnaire, are imputed by a "hot deck" method. In this procedure, a donor household is selected at random from a group of respondents that have similar characteristics.

Income and expenditure data are imputed by the nearest-neighbour method. This technique involves forming groups of similar households or individuals based on certain criteria (e.g., province of residence). Within those groups, each household requiring imputation (recipient) is matched to a household that has a complete questionnaire (donor) and resembles the other most closely with respect to certain characteristics (e.g., income, number of children, number of adults). The donor's data are imputed to the recipient as long as they satisfy the edit requirement for consistency with the data reported by the recipient.

The SHS collects information about various aspects of household budgets. Imputation is not done for the whole questionnaire but by sections that generally correspond to the questionnaire's sections, i.e., by groups of interrelated questions. This maximizes the number of potential donors in the sense that a household which leaves only one question unanswered, for example, can serve as a donor for those sections which it answered in full. This approach implies that one household could receive data from more than one donor. That possibility is minimized by the fact that we look for the household that exhibits the greatest possible similarity with respect to certain characteristics, which are often the same from section to section. It is important to note that all the questions in one section are imputed by the same donor, which preserves the relationships between questions.

5.3 Identification of usable data

The data for certain households whose questionnaires are at least partially complete may be rejected during processing. There are two main reasons for rejection. First, when a large portion of the income or expenditure questions is left unanswered, the questionnaire is deemed incomplete and is not used. Second, questionnaires are considered unusable if, following processing (editing for consistency and imputation of missing data where necessary), the difference

between receipts (income and other money received by the household) and disbursements (expenditure plus net change in assets and liabilities) reported by the household is greater than 20%.

Once identified, the usable data are weighted to produce estimates.

6. WEIGHTING AND ESTIMATION

Estimates are based on the premise that each household in the sample represents a certain number of households in the target population, as it was defined in section 2. Accordingly, each respondent household is assigned a survey weight, which indicates how many households in the population it represents. The survey weight is generally the product of three factors: the sampling weight, which incorporates data from the sample design; a non-response adjustment factor, which compensates for non-respondent households; and an adjustment factor that reflects characteristics from sources other than the survey. Also included in the calculation of the survey weight is an adjustment factor for influential data, though it affects very few households.

6.1 Sampling weight

A household's sampling weight is the inverse of its probability of being included in the sample. Since the SHS is a probability survey, every household in the target population has a known probability of being selected for the sample. For example, if a household's selection probability is 1 over 200, its weight will be 200.

For a given sample allocation, the sampling weight is determined by the sample design. The SHS uses the LFS sample design, which is self-weighting by stratum (i.e., the sampling weights set when the design is developed are equal within each stratum). If the sample design and sample allocation remained unchanged, the initial weights could be used. However, the stabilization and subsampling steps described in section 3.3 alter the initial selection probabilities. The LFS sampling weights are adjusted to reflect those changes.

Since the SHS sample is a subset of the six LFS rotation groups, the SHS sampling weights are determined by adjusting the LFS sampling weights in accordance with the number of rotation groups used. This factor may vary from stratum to stratum since the number of rotation groups selected in each stratum is based on the SHS's specific sample allocation requirements.

6.2 Non-response adjustment

In instances where the respondent has failed to answer only some of the questions, the missing data are imputed by the methods described in section 5.2. In cases of total non-response (i.e., where the household cannot be contacted, where household members refuse to respond, or where the data provided cannot be used), the weights are adjusted.

Adjustment of weights for non-response is based on the premise that responding households can be used to represent all households, both responding and non-responding. For the purposes of this adjustment, the sample is first divided into non-response classes defined so as to increase the chances that respondents will have characteristics similar to non-respondents.

The non-response classes correspond to different levels of urbanization in each province or territory except Quebec, Ontario and British Columbia; these three provinces are first divided into two or three subprovincial regions. The urbanization levels are generally as follows: the principal metropolitan area, urban areas with a population of 100,000 to 500,000, smaller urban areas, and rural or remote areas. In some regions or provinces, some levels may have to be combined because their samples are too small. In each territory, there are only two non-response classes: the principal city and the rest of the target population.

High-income household strata also form specific non-response classes in each province where they exist. Note that non-response areas do not overlap; when combined, they cover the entire target population.

For each non-response class, a non-response adjustment factor corresponding to the inverse of the class's weighted response rate is computed. Expressed another way, this factor is the ratio of the number of households sampled, multiplied by the sampling weights so that they represent the class's households, to the number of weighted respondent households. To ensure that the non-response adjustment factors are not excessive, some non-response classes are combined when the adjustment factor is greater than 2.

6.3 Adjustment using auxiliary information

In theory, estimates can be produced by multiplying the sampling weight by the non-response adjustment. However, the estimates can be improved with auxiliary data about the target population. If the auxiliary data are correlated with the principal characteristics measured by the survey, more reliable estimates can be produced. For example, a household's expenditures are correlated with its size. A poor sample allocation with respect to household size would have an impact on the expenditure estimates. On the basis of auxiliary data on the number of households by size, the weights can be adjusted to obtain the real distribution of the number of households.

In the SHS, various sources of auxiliary data are used to adjust the weights. First, postcensal estimates produced by Statistics Canada's Demography Division provide population counts by age group and sex for each province and territory. Those counts are population projections for a given period based on census data and information from administrative records, such as births, deaths, immigration and emigration. After the counts have been adjusted to reflect the SHS's target population at the end of the survey's reference year, the estimates for 18 different age-sex groups for each province[8] are used to adjust the weights.

[8] In the territories, only four groups are used: two age groups for each sex.

The 18-and-over and under-18 population counts for selected metropolitan areas[9] are also used.

Estimates of the number of households by size (one, two or three or more persons) for each province and territory and by selected household type are also used to adjust the sample's representativeness in those groups. For household type, we use specifically the number of households composed of lone-parent families and the number of households composed of parents with never-married children.

In order to remedy certain problems observed in the income distribution of survey respondents, the population counts for some income classes based on administrative sources are also used to adjust the SHS weights. The data are the numbers of individuals who earned wage and salary income as reported by employers. Six income classes are used. Their boundaries are based on the following percentiles of the distribution: 25, 50, 60, 75, 90 and 95.[10] Since the administrative data for the reference year are not available when weighting of the SHS is carried out, the class counts are projected on the basis of administrative data for the previous year and trends in the LFS distribution of individuals by salary.

Adjustment of the weights to reflect all of the above figures is carried out simultaneously using a variant of the generalized regression (GREG) estimator based on the weighting method proposed by Lemaître and Dufour [4]. This method allows concordance between the survey estimates and the estimates from auxiliary sources, while ensuring that, after adjustment, all members of the same household still have the same weight. The adjustment factor produced by the GREG estimator is then applied to the sampling weight and the non-response adjustment factor to generate the household's final weight.

6.4 Adjustment for influential data

Since expenditures have a highly asymmetric distribution, the samples used in expenditure surveys are prone to having extreme values. If a particular household has a combination of extreme values and a high weight, it may contribute disproportionately to the estimates. The presence of such influential data has a serious effect on estimates of totals and averages, chiefly at the provincial level and for subsets of the population.

To minimize the negative impact of these influential data on interprovincial comparisons and trend estimates, the weights of some households are adjusted to reduce their contribution to the estimates. To make such an adjustment, we use auxiliary information in the form of the distribution of individuals' incomes, based on their tax data. Since total expenditure is very closely tied to income, the correction will weaken the impact on estimates of total expenditures.

[9] St. John's, Halifax, Saint John, Quebec City, Montreal, Ottawa, Toronto, Winnipeg, Regina, Saskatoon, Calgary, Edmonton, Vancouver and Victoria.
[10] For some provinces, the 98th or 99th percentile is used instead of the 95th percentile.

The approach involves identifying the few individuals who make a major contribution (usually more than 1%) to the provincial estimates of total income. If necessary, we then adjust the weight of the individuals' households to ensure that the estimated number of individuals with that level of income does not exceed the number obtained from the distribution of individuals' tax data. Then the auxiliary information adjustment described earlier is applied again to ensure consistency.

It is important to note that the influential data adjustment, which focuses on extreme data, affects the weight of very few households (usually fewer than five for the whole sample).

6.5 Estimates

When the SHS estimates of average household expenditure are produced, households that existed for only part of the reference year, called part-year households, are excluded. (See section 4.1). Part-year households are composed entirely of persons who were members of other households for part of the survey year, as in the example of the two young adults living with their parents who get married and form a new household during the reference period. There are also households composed solely of persons who immigrated to Canada during the reference period. Part-year households make up a very small proportion of the household sample (less than 4%).

On the other hand, when estimates of the total expenditures of the Canadian population or a subpopulation are produced, the entire sample of households is used.

7. ESTIMATION OF SAMPLING ERROR

After the estimates have been computed, their reliability must be measured; in other words, the sampling error associated with each estimate must be estimated. The usual measure of sampling error is the standard error or the coefficient of variation (which is simply the standard error expressed as a percentage of the estimate). The standard error is the degree of variation observed in the estimates following the selection of one particular sample rather than another. Since the SHS is a probability survey, the standard error of its estimates can be estimated.

In the SHS, the jackknife method is used to estimate the standard error. This technique involves creating replicates of the sample based on SHS data. The same number of replicates is generated, as there are primary sampling units (PSUs), with one PSU being removed from the sample for each successive replicate. Each PSU belongs to a stratum, and when the PSU is removed, the sampling weights of the other PSUs in the stratum are adjusted accordingly. Then the final estimates are recomputed with the auxiliary data adjustments described in section 6.3 applied to the replicates. By repeating this operation for each PSU in the sample, we obtain as many estimates as there are PSUs. The variability of these estimates is used to estimate the standard error of the

estimate for the entire sample. The mathematical formula is shown in Appendix 1.

It is important to note that in the SHS, estimates of the standard error or coefficient of variation ignore the fact that some data were imputed. As a result, the computed CVs may underestimate the actual values. For most of the survey's variables, the effect of imputation is minimal. The impact of imputed data for each expenditure variable is included in the data quality report for each survey year.

7.1 Model for approximating the CV for domain estimates

For operational reasons, CVs cannot be produced for every characteristic collected by the survey at every level of aggregation of possible interest to users (e.g., by income quintile, household type, level of urbanization, tenure, selected metropolitan areas). The approach suggested to SHS users is to compute an approximate CV using a relationship between the number of households in the sample reporting expenditures for a category and the CV at an aggregate level (generally the national level). That relationship, based on the CV's tendency to grow in direct proportion to the decline in the square root of the number of households reporting a particular expenditure, is illustrated in Appendix 2.

7.2 Model for approximating the CV based on the microdata file

Microdata file users can take another approach to approximating the CVs of estimates. This approximation is generally more effective than the one described above. The method, which is fairly simple to use, is described in greater detail in reference [5]. It can be used only in combination with the microdata file, since the data and weights for all households are needed to compute the approximation.

8. DATA SUPPRESSION AND CONFIDENTIALITY

Steps are taken to ensure that the SHS estimates are sufficiently reliable to be published and that the anonymity of respondent households is maintained.

8.1 Suppression of unreliable data in estimate tables

Since the coefficient of variation is an indicator of data reliability, ideally we would use it to determine whether the estimates should be published or not. Estimates whose estimated CV is greater than 33% are not reliable enough to be released.

However, because so many estimates are produced for the SHS, it is impossible to compute the CV for each one. A study based on FAMEX data showed that the CVs generally reach 33% when the number of households reporting an expenditure approaches 30. This rule is used to determine whether SHS estimates can be published or not. Since it is a rule of thumb, some estimates will be released even if their CVs are above 33%, while others will not be published

even though their CVs are under 33%. An assessment of this rule's performance is included in the 1997 data quality report [3].

It should be noted that even if the estimates of average expenditures for a certain type of purchase are not disseminated because they were reported by fewer than 30 households, the data are reflected in the estimates for aggregate components.

8.2 Confidentiality of microdata files

Even though a public use microdata file is produced from SHS data, it is different from the one used by Statistics Canada for the release of estimates. The differences are largely due to a series of measures taken to protect the anonymity of the responding households.

9. CHANGES IN THE SURVEY METHODOLOGY

The introduction of the annual Survey of Household Spending has provided more frequent, more reliable estimates of expenditures, particularly at the provincial level since the total sample was increased and its allocation was revised. A new questionnaire, much less detailed than the one used in the Family Expenditure Survey, was developed for the 1997 SHS. Nevertheless, the survey's methodology has changed little from year to year, with the exception of the auxiliary data adjustment applied during weighting.

For the 1999 survey, population projections based on the 1996 Census took the place of the 1991 projections used in previous surveys. In addition, the weighting strategy was altered as part of a project to harmonize auxiliary data adjustments in Statistics Canada's income surveys. For the SHS, the major changes were the use of many more age-sex groups for population counts, and the introduction of counts for household types and for certain wage and salary classes. As a result of all these changes, a historical revision of estimates from the 1997 and 1998 SHS and the 1996 and 1992 FAMEX is being carried out to ensure comparability in trend analyses.

BIBLIOGRAPHY

[1] *Users' Guide for the Survey of Household Spending*, Statistics Canada (available for each survey year), Catalogue No. 62F0026MIE

[2] *Methodology of the Canadian Labour Force Survey*, Catalogue Number 71-526-XPB

[3] *1997 Survey of Household Spending – Data Quality Indicators*, Household Survey Methods Division (HSMD), Internal document, Statistics Canada (also available for each survey year), Catalogue No. 62F0026MIE

[4] Lemaître and Dufour (1987), *An Integrated Method for Weighting Persons and Families, Survey Methodology*, Vol.13, no 2, pp.211-220, Statistics Canada

[5] Beaumont, J.-F. (2000), *Variance Estimation For A Public Use Microdata File From A Complex Survey*, HSMD Working Paper, HSMD-2000-002F/A, Statistics Canada

APPENDIX 1

Formula for computation of the variance of estimates by the jackknife method

In the jackknife method of estimating the variance of estimates, the variability of the estimates is measured with the following formula:

$$Var(\hat{Y}) = \sum_{h=1}^{H} \frac{n(h)-1}{n(h)} \sum_{i=1}^{n(h)} (\hat{Y} - \hat{Y}_{(hi)})^2$$

where

n(h) is the number of PSUs in stratum h

$\hat{Y}_{(hi)}$ is the estimate of Y when PSU i is removed from stratum h.

The standard error is the square root of the variance.

APPENDIX 2

Formula for approximating the CV for a domain (a population subgroup)

If CV (Y) represents the CV for the estimate of the household average of a particular characteristic for the whole population, then we can compute an approximate CV for the estimate of that characteristic for a domain (which can be taken as a subgroup of the population, such as a household type, an income quintile or a level of urbanization) using the following equation:

$$CV\ (Y_d) = CV\ (Y) \times \sqrt{\frac{nP}{n_d P_d}}$$

where

n : *number of households in the sample*
P : *estimate of the proportion of households reporting a value > 0 for this characteristic in the population*
n_d : *number of households in the sample for domain d*
P_d : *estimate of the proportion of households reporting a value > 0 for this characteristic in domain d*

The CV, size n and proportion P for the national level are generally used to calculate the approximations for the various domains. Where we wish to compute an approximate CV for a metropolitan area, we can use provincial values since the domain is entirely within a single province and since provincial CVs will be published for the SHS.

Division de la statistique du revenu

62F0026MIF - 01003

La méthodologie de l'enquête sur les dépenses des ménages

Préparé par :
Sophie Arsenault
Johanne Tremblay

Octobre 2001

Statistique Canada Statistics Canada

Canada

Des données sous plusieurs formes

Statistique Canada diffuse les données sous formes diverses. Outre les publications, des totalisations habituelles et spéciales sont offertes. Les données sont disponibles sur Internet, disque compact, disquette, imprimé d'ordinateur, microfiche et microfilm, et bande magnétique. Des cartes et d'autres documents de référence géographiques sont disponibles pour certaines sortes de données. L'accès direct à des données agrégées est possible par le truchement de CANSIM, la base de données ordinolingue et le système d'extraction de Statistique Canada.

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Services aux clients, Division de la statistique du revenu, Statistique Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355: revenu@statcan.ca) ou à l'un des centres de consultation régionaux de Statistique Canada :

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

Vous pouvez également visiter notre site sur le Web :http://www.statcan.ca

Un service d'appel interurbain sans frais est offert **à tous les utilisateurs qui habitent à l'extérieur des zones de communication locale** des centres de consultation régionaux.

Service national de renseignements	**1 800 263-1136**
Service national d'appareils detélécommunications pour les malentendants	**1 800 363-7629**
Numéro pour commander seulement (Canada et États-Unis)	**1 800 267-6677**

Renseignements sur les commandes et les abonnements

Les prix ne comprennent pas les taxes de vente

On peut se procurer ce produit nº 62F0026MIF-01003 au catalogue sur internet gratuitement. Pour obtenir un numéro de ce produit, les utilisateurs sont priés de se rendre à http://www.statcan.ca/cgi-bin/downpub/research_f.cgi.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois et dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec le centre de consultation régional de Statistique Canada le plus près de chez vous.

Statistique Canada
Division de la statistique du revenu

Méthodologie de l'enquête sur les dépenses des ménages

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2001

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

octobre 2001

N° 62F0026MIF - 01003 au catalogue

Périodicité : Irrégulier

Ottawa

This publication is available in English upon request.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Résumé

L'objectif de ce document est de fournir une description détaillée de la méthodologie de l'Enquête sur les dépenses des ménages. Les sujets traités incluent: la population cible; le plan d'échantillonnage; la collecte des données; le traitement des données; la pondération et l'estimation; l'estimation de l'erreur d'échantillonnage; et la suppression des données et la confidentialité.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

TABLE DES MATIÈRES

1. INTRODUCTION **9**

2. LA POPULATION CIBLE **10**

3. LE PLAN D'ÉCHANTILLONNAGE **10**
3.1 La taille et la répartition de l'échantillon de l'EDM 11
3.2 Le plan d'échantillonnage de l'EPA (la sélection des grappes) 12
3.3 La sélection de l'échantillon de l'EDM 14
3.4 Les particularités du plan dans les territoires 15

4. LA COLLECTE DES DONNÉES **16**
4.1 Méthode de collecte des données 16
4.2 L'entrevue et les procédures de suivis 17
4.3 La surveillance et les contrôles 18
4.4 La non-réponse à l'EDM 19

5. LE TRAITEMENT DES DONNÉES **19**
5.1 Le codage et la saisie des données 19
5.2 La vérification et l'imputation des données 20
5.3 L'identification des données utilisables 21

6. LA PONDÉRATION ET L'ESTIMATION **21**
6.1 Le poids de sondage 21
6.2 L'ajustement pour la non-réponse 22
6.3 L'ajustement à l'information auxiliaire 23
6.4 L'ajustement pour les données influentes 24
6.5 Les estimations 25

7. L'ESTIMATION DE L'ERREUR D'ÉCHANTILLONNAGE **25**
7.1 Modèle pour dériver une approximation du CV pour les estimations des domaines 26
7.2 Modèle pour dériver une approximation du CV à partir du fichier de microdonnées 26

8. LA SUPPRESSION DE DONNÉES ET LA CONFIDENTIALITÉ **26**
8.1 La suppression des données non fiables dans les tableaux d'estimations 26
8.2 La confidentialité des fichiers de microdonnées 27

9. LES CHANGEMENTS DANS LA MÉTHODOLOGIE DE L'ENQUÊTE **27**

BIBLIOGRAPHIE **28**

ANNEXE 1
Formule pour le calcul de la variance des estimations à partir du jackknife 29

ANNEXE 2
Formule d'approximation du CV pour un domaine (un sous-groupe de la population) 30

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

1. INTRODUCTION

L'enquête sur les dépenses des ménages (EDM) est une enquête annuelle recueillant des informations auprès des ménages canadiens sur les habitudes de dépense, les caractéristiques des logements et l'équipement ménager.

L'EDM a été instaurée en janvier 1998 pour la collecte des dépenses des ménages de l'année 1997. Cette enquête remplace l'enquête périodique sur les dépenses des familles (EDF) qui avait généralement lieu à tous les 4 ans.[1] L'enquête annuelle a été instaurée pour répondre à un besoin de données provinciales plus précises et plus fréquentes pour les comptes nationaux. Par conséquent, elle bénéficie d'un échantillon plus grand. Un autre changement majeur implanté avec la nouvelle enquête fut un remaniement complet du questionnaire. Le niveau de détail demandé pour les dépenses a été beaucoup réduit ce qui s'est traduit par une diminution du temps d'entrevue de près de 33%.

Les objectifs de l'EDM sont nombreux puisque cette enquête est une source de données précieuse pour plusieurs produits de Statistique Canada ainsi que pour un grand nombre d'utilisateurs à l'extérieur de Statistique Canada. Outre les comptes nationaux qui utilisent les estimations de l'EDM dans le développement de leurs estimations en dépenses personnelles de biens et services au niveau national et provincial, les données de l'EDM sont la source d'information pour la mise à jour du panier de consommation utilisé dans le calcul de l'indice des prix à la consommation. Elles servent aussi à déterminer les seuils de faible revenu, utilisés par l'enquête sur la dynamique du travail et du revenu pour établir la proportion d'individus ou de familles à faible revenu. De plus, les données de l'EDM correspondent à une des cinq sources de données pour la création du modèle de simulation des politiques sociales qui permet l'analyse de l'impact de diverses politiques économiques et sociales. Depuis l'abolition de l'enquête sur l'équipement ménager, l'EDM est devenue la source des données sur les caractéristiques du logement et l'équipement des ménages nécessaires à la Société Canadienne d'hypothèque et de logements ainsi qu'à un grand nombre d'utilisateurs externes.

Les données de l'EDM sont recueillies à partir d'entrevues personnelles réalisées auprès d'un échantillon de ménages des 10 provinces et des 3 territoires canadiens. Les ménages sont contactés au début de l'année, entre janvier et mars, et doivent rapporter les dépenses qu'ils ont encourues au cours de l'année civile précédente. Les données sont par la suite vérifiées et pondérées. Les produits de l'enquête comprennent les tableaux et fichiers de microdonnées nécessaires aux divers utilisateurs.

L'objectif de ce document est de fournir une description détaillée de la méthodologie de l'enquête quant au plan d'échantillonnage, à la collecte et au traitement des données, à la production des estimations et des autres produits

[1] La dernière enquête sur les dépenses des familles portait sur les données de l'année 1996. Des données au niveau national avaient également été recueillies pour les années 1992, 1986, 1982 et 1978. Pour certaines années, par exemple 1990 et 1984, l'enquête a été menée dans certaines grandes villes seulement.

ainsi qu'aux règles régissant leur diffusion. Pour une description plus sommaire de la méthodologie, il est préférable de se référer au manuel de l'utilisateur disponible pour chacune des années d'enquête [1].

2. LA POPULATION CIBLE

La population cible de l'EDM comprend les individus vivant dans les ménages privés canadiens à l'exclusion des représentants officiels de pays étrangers qui vivent au Canada et leurs familles ainsi que les résidents des réserves indiennes et des terres publiques. La restriction aux ménages privés implique que les pensionnaires d'établissements institutionnels tels que les prisons, les hôpitaux pour malades chroniques, les résidences pour personnes âgées ainsi que les membres d'ordres religieux et d'autres groupes vivant en communauté, les membres des Forces Armées vivant dans les camps militaires et les individus vivant de façon permanente dans les hôtels ou les maisons de chambres sont exclus.

L'enquête couvre près de 98% de la population dans les 10 provinces. Au Yukon, aux Territoires du Nord-Ouest et au Nunavut, la couverture correspond respectivement à 81%, 92% et 89% de la population (ou 80%, 93% et 90% des ménages) puisque les personnes vivant dans les très petites communautés (généralement constituées de moins de 100 ménages) ou dans les régions non organisées sont également exclues.

Il est important de noter que jusqu'en 1996, l'enquête périodique sur les dépenses des familles ne couvrait que les villes de Whitehorse et Yellowknife dans les territoires. Pour l'enquête sur les dépenses des ménages de 1997, la couverture a été étendue à 78% de la population du Yukon et 70% des Territoires du Nord-Ouest et du Nunavut pour atteindre la couverture actuelle avec l'enquête de 1998. Depuis l'EDM 2000, les territoires ne sont enquêtés qu'une année sur deux de sorte à réduire le fardeau de réponse.

3. LE PLAN D'ÉCHANTILLONNAGE

Les données sur les dépenses sont recueillies auprès d'un échantillon de ménages sélectionné selon un plan d'échantillonnage stratifié à plusieurs degrés. Le plan varie selon le niveau d'urbanisation mais généralement il s'agit d'un échantillon à deux degrés dont le premier degré est un échantillon aréolaire c'est-à-dire un échantillon d'aires géographiques que l'on désignera des grappes. La liste de tous les logements privés se trouvant dans les grappes sélectionnées est ensuite établie pour permettre la sélection d'un échantillon de logements qui constitue le second degré d'échantillonnage. Tous les logements choisis qui sont habités par des individus de la population cible constituent l'échantillon de l'enquête.

L'EDM utilise principalement le même plan d'échantillonnage que l'enquête sur la population active (EPA) dans le but de minimiser les coûts d'opération. Les logements de l'échantillon de l'EDM sont choisis dans les grappes

échantillonnées de l'EPA mais l'échantillon de logements de l'EDM est différent de celui de l'EPA. Les grandes lignes du plan de l'EPA pour la sélection des grappes sont décrites dans la section 3.2. Une description plus détaillée peut être obtenue en consultant la publication de la méthodologie de l'EPA [2]. Les caractéristiques particulières à la sélection des logements de l'EDM à l'intérieur des grappes échantillonnées de l'EPA sont ensuite décrites dans la section 3.3. Le plan d'échantillonnage dans les territoires est différent. Les particularités du plan dans ces régions sont présentées dans la section 3.4. La section qui suit explique d'abord comment l'échantillon est distribué à travers les provinces et les territoires.

3.1 La taille et la répartition de l'échantillon de l'EDM

La taille de l'échantillon de l'EDM peut varier selon les années. Suite au lancement de l'enquête annuelle de 1997, la taille d'échantillon a été établi à près de 24000 ménages, soit une augmentation d'environ 67% par rapport à l'enquête périodique sur les dépenses des familles. Au cours des années qui ont suivi, la taille d'échantillon a varié légèrement en fonction des contraintes budgétaires et de la présence ou non des territoires dans l'enquête. La taille totale d'échantillon pour chaque année est indiquée dans le tableau 3.1.

À chaque année, l'échantillon total est réparti entre les provinces et les territoires (s'il y a lieu) de sorte à obtenir des estimations avec un niveau de fiabilité similaire. Plus précisément, la répartition est effectuée en fonction de la variabilité du revenu dans chaque province avec une plus grande portion de l'échantillon allouée aux provinces pour lesquelles les revenus des individus diffèrent plus les uns des autres. On tient également compte, mais avec une importance beaucoup moindre, de la taille de la population dans la répartition de l'échantillon. Pour les territoires et pour l'IPE où la population est beaucoup plus petite que dans les autres provinces, la taille d'échantillon est déterminée au préalable pour éviter qu'une trop forte proportion de la population se retrouve dans l'échantillon. Actuellement, l'EDM échantillonne environ 4% de la population dans chacun des territoires et 2% de la population de l'IPE.

Les taux de réponses observés dans les enquêtes antérieures servent à ajuster les tailles d'échantillon de chaque province ou territoire. Il en est de même pour les taux de vacance (la proportion de logements inoccupés) au niveau provincial et infra-provincial. Ces derniers proviennent des données les plus récentes de l'EPA pour la période de janvier à mars correspondant à la période de collecte de l'EDM.

L'échantillon provincial ou territorial est finalement distribué proportionnellement à la taille de la population dans les régions métropolitaines de recensement[2]. La portion de l'échantillon allouée à l'extérieur des régions métropolitaines de recensement suit la répartition de l'échantillon de l'enquête sur la population active [2] sauf aux T.N.-O. et au Nunavut où l'EPA n'a pas lieu.

Les tailles d'échantillon provinciales et territoriales des enquêtes sur les dépenses des ménages depuis 1997 sont présentées dans le tableau 3.1 en

[2] Ainsi que pour les villes de Charlottetown, Summerside, Whitehorse, Yellowknife et Iqaluit.

terme du nombre de ménages c'est-à-dire après avoir exclu les logements sélectionnés qui étaient vacants ou inéligibles. Dans l'enquête de 1997, une plus grande proportion de l'échantillon a été allouée aux provinces de Terre-Neuve, de la N.É. et du N.B. parce qu'on prévoyait que la taille de l'échantillon augmenterait dans les années subséquentes et on voulait attribuer immédiatement l'échantillon augmenté aux provinces qui avaient signé l'entente pour l'harmonisation de la taxe sur les produits et services. Par la suite, il s'est avéré que le budget permettrait tout au plus une taille d'échantillon de 24 000 ménages et la répartition de l'échantillon a été corrigée pour les années subséquentes.

Tableau 3.1
Taille d'échantillon (nombre de ménages) par province ou territoire

Provinces et territoires	Taille d'échantillon (nombre de ménages)			
	EDM 1997	EDM 1998	EDM 1999	EDM 2000[3]
Canada	**23 842**	**20 236**	**23 518**	**20 877**
Terre-Neuve	1 997	1 343	1 937	1 794
Î.-P.-É.	795	807	822	822
N.-É.	2 424	1 573	2 199	2 040
N.-B.	2 044	1 406	1 957	1 821
Québec	3 122	2 848	2 710	2 516
Ontario	3 362	3 056	3 453	3 202
Manitoba	1 772	1 739	2 034	1 882
Saskatchewan	1 478	1 721	1 837	1 697
Alberta	2 743	2 186	2 519	2 336
C.-B.	3 010	2 590	2 985	2 768
Total des provinces	**22 747**	**19 269**	**22 453**	**20 877**
Yukon	451	383	403	0
T.N.-O.	644	383	414	0
Nunavut		201	248	0
Total des territoires	**1 095**	**967**	**1 065**	**0**

3.2 Le plan d'échantillonnage de l'EPA (la sélection des grappes)

Le plan d'échantillonnage de l'EPA est élaboré à partir des données du recensement de la population et remanié après chaque recensement décennal pour tenir compte des changements de la population. Le plan actuel est basé sur les données du recensement de 1991.

Les principes du plan d'échantillonnage de l'EPA sont les mêmes dans chaque province. D'abord chacune d'elles est divisée en plusieurs régions géographiques qui correspondent aux intersections des régions économiques

[3] Pour 2000, il s'agit d'une approximation du nombre de ménages dans l'échantillon basé sur le taux de vacance et de ménages éligibles de l'enquête précédente. Le nombre de ménages dans l'échantillon n'est déterminé exactement qu'après que les logements sélectionnés aient été visités par les intervieweurs pour exclure les logements vacants ou habités par des individus non éligibles à l'enquête.

(RÉ) et des régions économiques d'assurance-emploi (RÉAE). Entre autres, chaque région métropolitaine de recensement constitue une région géographique puisqu'elle correspond à une RÉAE.

Chacune de ces régions géographiques est ensuite subdivisée en types de secteurs. On y retrouve principalement les secteurs urbains, les secteurs ruraux et les régions éloignées. Le plan d'échantillonnage varie selon le type de secteur.

Les secteurs urbains

Dans certaines grandes villes où le nombre d'immeubles à appartements est élevé, on utilise à la fois une base de sondage d'appartements et une base aréolaire alors que pour les autres secteurs urbains on utilise uniquement la base aréolaire.

La base aréolaire est une liste d'aires géographiques représentant chaque secteur. Des strates sont formées en combinant ces aires géographiques. On peut avoir jusqu'à trois niveaux de stratification. Généralement les premiers niveaux de stratification visent à former des strates géographiquement compactes et contiguës alors que le dernier niveau permet la création de strates finales aussi homogènes que possible en fonction de certaines caractéristiques socio-économiques. Dans quelques grandes villes[4], on forme des strates distinctes à partir des secteurs de dénombrement affichant un revenu moyen des ménages élevé (environ 100 000$ ou plus).

Pour réduire les coûts de collecte, on ne sélectionne pas directement les ménages qui font partie de la strate finale. On divise plutôt la strate en grappes. Dans les secteurs urbains, les grappes peuvent être des combinaisons de cotés d'îlots, des secteurs de dénombrement (SD) ou des sections de SD. On choisit ensuite des grappes (généralement six, parfois un multiple de six) dans chacune des strates avec une probabilité proportionnelle à la taille de la grappe. Ainsi si une grappe est deux fois plus grande qu'une autre, elle aura deux fois plus de chances d'être choisie que la seconde.

La base de sondage d'appartements est une liste d'appartements créée à partir d'information provenant de la Société Canadienne d'hypothèque et de logements. Cette liste permet une meilleure représentativité des locataires d'appartements et minimise l'effet de croissance importante des grappes que l'on retrouve avec une base aréolaire lorsque de nouveaux immeubles à appartements sont construits dans la grappe. Dans certaines de ces villes[5], les appartements sont divisés en deux groupes : les strates d'appartements à bas revenu (où le revenu moyen des ménages qui y habitent est inférieur à 20 000$) et les strates régulières. Pour chaque strate d'appartements de cette base, on choisit des immeubles à appartements comme échantillon de premier degré avec une probabilité proportionnelle au nombre d'appartements dans l'immeuble.

---

[4] On retrouve ce type de strates à Montréal, Ottawa, Toronto, Hamilton, London, Winnipeg, Calgary et Vancouver.

[5] Les villes de Montréal, Ottawa-Hull, Toronto, Winnipeg, Calgary, Edmonton et Vancouver.

Dans les secteurs urbains à faible densité, qui sont de petites villes très dispersées, on utilise un plan d'échantillonnage différent. Il s'agit d'un plan à trois degrés où on choisit d'abord des villes dans les strates, ensuite des grappes (coté d'îlots) dans les villes et finalement des logements dans les grappes.

Les secteurs ruraux

Dans les secteurs ruraux, on utilise exclusivement une base aréolaire. On forme des strates géographiques en regroupant deux à trois divisions de recensement que l'on subdivise, lorsque le nombre le permet, pour former des strates homogènes par rapport à des caractéristiques socio-économiques. Au premier degré d'échantillonnage, on choisit des secteurs de dénombrements dans chacune des strates finales avec une probabilité proportionnelle au nombre de ménages du SD.

Dans les secteurs ruraux avec une faible densité de population, on applique une variante au plan d'échantillonnage. On choisit deux ou trois unités primaires constituées d'un groupe de six SD, au premier degré, et par la suite on sélectionne un échantillon de logements dans chacun des SD des unités primaires choisies.

Les régions éloignées

La portion nordique des provinces (excluant les maritimes) est, en grande partie, peu peuplée. Généralement l'échantillon est sélectionné en deux étapes. On sélectionne d'abord un échantillon d'agglomérations que l'on appelle lieux, et de SD. Dans une région éloignée du Québec, on a également recours à un échantillonnage à trois degrés.

Les lieux comptant moins de 10 ménages ou 25 personnes sont omis du plan ainsi que les SD comptant moins de 25 ménages. Malgré ces omissions, le plan couvre 90% de la population des régions éloignées des provinces.

3.3 La sélection de l'échantillon de l'EDM

Toutes les grappes sélectionnées dans le plan de l'EPA sont visitées par les intervieweurs et ceux-ci produisent une liste de tous les logements privés s'y trouvant. À partir de cette liste, un échantillon de logements est choisi pour l'EPA et un échantillon différent est choisi pour l'EDM. Les logements sont choisis à partir d'un échantillon systématique.

Comme la taille d'échantillon de l'EDM est beaucoup plus petite que celle de l'EPA, on ne choisit pas des logements dans toutes les grappes sélectionnées de l'EPA. L'EPA est une enquête par panel où les ménages demeurent dans l'échantillon pendant six mois. Le plan de l'EPA a été établi de sorte à ce qu'il puisse être divisé en six sous-échantillons représentatifs pour permettre le renouvellement chaque mois du sixième de l'échantillon. C'est d'ailleurs pourquoi on sélectionne six grappes (ou un multiple de 6) dans chacune des strates finales, une par groupe de renouvellement. Cette approche permet de sélectionner facilement un échantillon de taille plus petite pour une autre enquête

puisqu'il est possible de choisir un sous-ensemble des groupes de renouvellement. Les enquêtes supplémentaires de l'EPA utilisent généralement cette approche en prenant un sous-ensemble des 6 groupes de renouvellement. Dans le cas de l'EDM, le nombre de groupes de renouvellements à utiliser est déterminé au niveau des strates pour répondre aux besoins spécifiques de l'enquête en terme de la répartition provinciale et infra-provinciale de l'échantillon. Il arrive que seulement une portion d'un groupe de renouvellement soit nécessaire. Dans ce cas, on enlève des ménages au hasard.

L'échantillon de logements est obtenu suite aux opérations de listage des grappes. Comme les taux d'échantillonnage sont déterminés au préalable, il peut y avoir un écart entre la taille d'échantillon prévue et la taille d'échantillon obtenue si le nombre de logements sur la liste diffère de celui utilisé lors du développement du plan de sondage de l'enquête. Pour éviter une augmentation des coûts de collecte (puisque la taille des grappes est surtout portée à croître) ainsi qu'une grande disparité dans la charge de travail des intervieweurs, on a recours à deux méthodes de contrôle de la taille d'échantillon.

On corrige généralement le problème en supprimant certains logements sélectionnés originalement de façon aléatoire. Cette opération qui permet de garder la taille d'échantillon au niveau voulu est appelée la stabilisation de l'échantillon. Dans le cas d'expansions importantes du nombre de logements dans certains secteurs urbains, on procède plutôt à un sous-échantillonnage des grappes. Selon l'importance de l'expansion du nombre de logements dans la grappe et la similarité de ces nouveaux logements avec les autres de la même strate, on formera des sous-grappes, on créera une nouvelle strate ou on sous-échantillonnera les logements de la grappe.

3.4 Les particularités du plan dans les territoires

Comme l'enquête sur les dépenses des ménages doit couvrir l'ensemble des territoires alors que l'enquête périodique sur les dépenses des familles ne couvrait que Whitehorse et Yellowknife, un nouveau plan d'échantillonnage a été implanté pour les territoires lors de l'enquête de 1998[6]. Ce plan a été conçu différemment pour tenir compte du fait qu'une importante partie de la population est dispersée dans des communautés à faible densité. Cette particularité a un impact important sur les coûts de collecte de l'enquête. Ainsi, les individus vivant dans les régions non organisées, dans de très petites communautés (généralement de moins de 100 ménages) ou encore dans des secteurs difficilement accessibles ont été exclus de la population cible de l'enquête.

Nonobstant cette différence au niveau de la couverture qui fait en sorte qu'environ 19% de la population du Yukon est exclue de l'enquête, l'approche utilisée dans ce territoire est similaire à celle des provinces puisque l'EPA est menée également au Yukon.

[6] L'année 1997 fut une année de transition, le plan de sélection probabiliste dans les territoires n'étant pas encore finalisé. À partir d'un choix arbitraire de communautés ainsi que de l'échantillon des villes de Whitehorse et Yellowknife, on a été en mesure de fournir des estimations représentant 78% de la population du Yukon et 70% des Territoires du Nord-Ouest et du Nunavut combinés.

Aux Territoires du Nord-Ouest et au Nunavut, un plan spécifique à l'EDM a été instauré puisque l'EPA ne collectait pas d'information dans ces territoires. Ce plan est basé sur les données du recensement de 1996 alors que pour le Yukon et les provinces le plan de l'EPA est basé sur le recensement de 1991.

Les villes de Yellowknife et Iqaluit forment chacune une strate distincte divisée en grappes avec un échantillon de logements choisi dans chaque grappe. Les autres communautés sont groupées en deux ou trois strates en fonction de caractéristiques socio-démographiques telles que la taille de la population, la proportion d'autochtones et le revenu moyen des ménages. Chaque communauté correspond à une grappe et un échantillon de deux ou trois grappes est choisi dans chacune de ces strates. Par la suite, un échantillon d'environ 30 logements est sélectionné dans chacune des grappes choisies.

Les données de l'EDM pour l'année de référence 2000 n'ont pas été collectées dans les trois territoires suite à la décision de collecter cette information seulement à tous les deux ans dans les territoires pour pallier à un fardeau de réponse élevé pour la population de ces régions.

4. LA COLLECTE DES DONNÉES

L'EDM vise à recueillir des renseignements sur le budget complet des ménages canadiens sur une base volontaire. Ces renseignements incluent les dépenses, les revenus ainsi que les variations de l'avoir et des dettes, sur une période de douze mois qui va du 1er janvier au 31 décembre de l'année de référence.

L'EDM collecte également des renseignements sur les caractéristiques des logements et sur l'équipement ménager que possèdent les ménages. Cette information doit refléter la situation prévalant en date du 31 décembre de l'année de référence.

4.1 Méthode de collecte des données

La collecte des données est effectuée au moyen d'entrevues en personne entre intervieweurs et répondants. Ces entrevues se déroulent au cours des mois de janvier à mars qui suivent l'année de référence de l'enquête.

L'EDM est une enquête-mémoire c'est-à-dire que le répondant doit se rappeler des dépenses qu'il a effectuées au cours de la période de référence d'un an. On encourage donc les répondants à consulter des documents se rapportant à la période de référence tels que les relevés d'hypothèque, les chéquiers, les états de comptes de carte de crédits, les rapports d'impôt, dans le but de réduire l'effort de rappel et de fournir des renseignements plus précis. Pour les articles achetés à intervalles réguliers, les dépenses sont collectées le plus souvent en demandant au répondant la quantité achetée, la fréquence des achats et le prix généralement payé pour pouvoir dériver une estimation des dépenses annuelles du ménage. Plus spécifiquement, les dépenses alimentaires sont habituellement collectées sur une courte période (une semaine ou un mois), pour ensuite être

ramenées à une base annuelle. L'EDM ne collecte que le total des dépenses alimentaires, les dépenses détaillées sont collectées à tous les quatre ans au moyen d'un carnet de dépenses dans l'enquête sur les dépenses alimentaires.

Le questionnaire de l'EDM recueille des renseignements au niveau du ménage comme les dépenses de logements et d'ameublement, les dépenses alimentaires, les dépenses de transport ou de loisirs. D'autres renseignements doivent être fournis individuellement pour les membres du ménage comme le revenu personnel, les impôts et les dépenses vestimentaires. Ces renseignements sont souvent obtenus par procuration.

Les membres d'un ménage

Pour obtenir les dépenses relatives à un ménage, il est essentiel de bien identifier les membres qui en font partie. La personne ou le groupe de personnes qui occupe un logement constitue un ménage. Pour l'EDM, les membres de ce ménage sont définis de la façon suivante:

i) toutes les personnes qui vivent dans le logement au moment de l'entrevue qui n'ont pas une résidence permanente ailleurs et ne sont pas membres d'un autre ménage;
ii) toute personne qui fut membre de ce ménage pendant la période de référence, ou une partie de celle-ci, même s'ils n'y vivent pas au moment de l'entrevue.

En rapportant les dépenses et le revenu du ménage, il est donc important d'inclure les dépenses et le revenu des membres qui ont quitté le ménage ou de ceux qui se sont joint au ménage au cours de l'année de référence[7]. Dans ce cas, les données doivent refléter la portion de l'année où l'individu était membre du ménage.

Il est également possible qu'un ménage ait existé seulement pendant une partie de l'année de référence. C'est le cas par exemple de deux jeunes adultes qui vivent chez leurs parents et qui se marient et constituent un nouveau ménage pendant la période de référence. Les dépenses de ces ménages couvrent seulement une partie de l'année de référence. Ces ménages sont donc traités de façon particulière dans le calcul de certaines estimations. Ce point est élaboré plus en détail à la section 5.5.

4.2 L'entrevue et les procédures de suivis

Les entrevues sont menées par des intervieweurs de Statistique Canada souvent également en charge de collecter l'information de l'EPA. Ces intervieweurs reçoivent une formation particulière à l'EDM.

[7] Pour les personnes qui se sont joint à un ménage, il est nécessaire de déterminer s'ils vivaient auparavant dans un ménage qui n'existe plus maintenant. Dans l'affirmative, cet ancien ménage n'a aucune chance d'être sélectionné. Les données sont collectées pour la partie de la période de référence précédant le changement de ménage sur un questionnaire différent.

La semaine précédant sa visite, l'intervieweur envoie par la poste une lettre de présentation aux occupants des logements sélectionnés soulignant l'importance de l'enquête et la confidentialité des renseignements recueillis. Il rend ensuite visite au ménage pour mener l'entrevue. Lorsque le moment de la visite est mal choisi, ce dernier fixe un rendez-vous pour revenir à un moment plus opportun. Lorsqu'il n'y a personne à la maison, de nombreuses autres tentatives sont faites pour contacter le ménage, par exemple faire des visites à différentes heures de la journée ou consulter un annuaire par numéro (reverse directories) pour obtenir un numéro de téléphone.

Étant donné la grande variété de renseignements à obtenir, la collecte peut parfois nécessiter de longues entrevues et il peut être nécessaire d'effectuer plus d'une visite pour obtenir les renseignements complets. En moyenne, il faut environ une heure quarante pour compléter l'interview. À la fin de l'entrevue, les répondants peuvent conserver un sommaire des dépenses qu'ils ont rapportées pour leurs propres dossiers.

Si une personne refuse de participer à l'EDM, le bureau régional envoie une lettre au logement pour souligner l'importance de l'enquête et de la coopération du ménage. Ensuite, l'intervieweur fait une deuxième visite (ou un appel). Si l'intervieweur est incapable d'obtenir la participation du ménage, il complètera un rapport de non-interview. Selon les commentaires fournis, l'intervieweur principal décidera s'il poursuit ou non les tentatives de conversion de refus.

4.3 La surveillance et les contrôles

Tous les intervieweurs de l'EDM travaillent sous la surveillance d'un intervieweur principal qui est chargé de veiller à ce que les intervieweurs connaissent bien les concepts et les méthodes de l'enquête ainsi que de contrôler périodiquement le travail et de revoir les documents remplis. Les intervieweurs principaux travaillent à leur tour sous la surveillance des gestionnaires de programme qui sont postés à chacun des bureaux régionaux de Statistique Canada.

La première vérification des données est effectuée par l'intervieweur qui s'assure que l'information est complète pour chacune des sections du questionnaire. Les questionnaires sont par la suite vérifiés par les intervieweurs principaux.

Comme la capacité de rappel des répondants est une composante essentielle à la qualité des données de l'EDM, une des mesures de contrôle consiste à mesurer la différence entre les rentrées d'argent (revenus et autres montants reçus par le ménage) et les sorties d'argent (dépenses plus la variation nette de l'actif et du passif) déclarés par le ménage. Si la différence est supérieure à 10% des rentrées ou 10% des sorties, selon le plus élevé des deux montants, l'intervieweur ou l'intervieweur principal communiquera à nouveau avec les répondants pour obtenir des renseignements supplémentaires et tenter d'identifier les erreurs ou les omissions.

4.4 La non-réponse à l'EDM

Malgré toutes les tentatives effectuées pour obtenir l'information, il reste toujours un certain nombre de ménages non répondants. Par exemple, il a été impossible de les contacter ; une entrevue n'a pas été conduite pour des circonstances incontrôlables; ou parce que les membres du ménage ont refusé de participer à l'enquête. Pour chacune des années d'enquête, on retrouve de l'information détaillée sur les taux de non-réponses dans le document sur la qualité des données[3]. Les taux de non-réponses à la collecte qui ont été observés au cours des dernières années sont présentés dans le tableau 4.1. On trouve également dans ce tableau le taux de non-réponse final qui tient compte des ménages exclus après avoir effectué le traitement des données (Voir 5.4).

Tableau 4.1
Taux de non-réponse à l'EDM

Année de référence	Taux de non-réponse à la collecte			Taux de non-réponse final (à l'estimation)
	TOTAL	Pas de contact	Refus	
1997	20.7	5.8	15.0	24.4
1998	20.7	4.9	15.8	23.6
1999	23.6	5.9	17.7	26.8

5. LE TRAITEMENT DES DONNÉES

Les principales étapes du traitement des données de l'EDM sont le codage des réponses, la saisie des données, la vérification, l'imputation des non-réponses partielles, l'identification des données utilisables et la pondération. La pondération sera traitée dans la section 6.

5.1 Le codage et la saisie des données

Dans l'EDM, les questions qui nécessitent un codage sont très peu nombreuses. Cette étape est effectuée par l'intervieweur, puis vérifiée par l'intervieweur principal. Les questionnaires sont ensuite groupés en lots de 20 questionnaires et les données inscrites dans les questionnaires sont saisies dans les bureaux régionaux de Statistique Canada. La saisie des données est vérifiée en sélectionnant un échantillon de questionnaires provenant de chaque préposé à la saisie qui seront saisis une seconde fois. Si le nombre d'erreurs dépasse un certain seuil pour un questionnaire, le lot complet est soumis à la saisie à nouveau. La taille de l'échantillon soumis à la vérification dépend du rendement passé des préposés à la saisie.

5.2 La vérification et l'imputation des données

Une première étape de vérification automatisée des questionnaires est effectuée après que chacun d'eux ait été vérifié manuellement par l'intervieweur et l'intervieweur principal. Plusieurs règles de cohérence essentielles entre les réponses rapportées sur le questionnaire doivent être respectées. On identifie également les situations inhabituelles qui pourraient justifier des corrections. Cette étape de vérification automatisée est effectuée dans les bureaux régionaux de Statistique Canada, ce qui permet de communiquer avec les répondants lorsque des renseignements supplémentaires sont nécessaires pour résoudre des incohérences dans les réponses qu'ils ont fournies. Les problèmes identifiés au cours de cette vérification sont résolus par les membres des équipes de résolution des erreurs relatives aux questionnaires spécialement formés. Par la suite, les données sont transmises au bureau central où on effectue d'autres vérifications des données et on corrige les réponses invalides.

Lorsque le répondant a omis de répondre à certaines questions seulement, on parle de non-réponse partielle et dans ce cas, les données manquantes sont imputées. L'approche pour imputer les données diffère selon qu'il s'agit de données catégoriques ou continues. Les données catégoriques prennent seulement des valeurs spécifiques (comme les questions auxquelles on répond par oui ou non et les questions sur le type de logement habité) alors que les données continues peuvent prendre n'importe quelle valeur numérique (comme les revenus et les dépenses).

Les données catégoriques que l'on trouve principalement dans les sections sur les caractéristiques du logement et l'équipement du logement, sont imputées à l'aide d'une technique "hot deck" où un ménage donneur est choisi de façon aléatoire parmi un groupe de ménages répondants possédant des caractéristiques semblables.

Les données sur le revenu et les dépenses sont imputées au moyen de la technique du plus proche voisin. Cette méthode consiste à créer des groupes de ménages ou d'individus semblables selon certains critères (par exemple, la province de résidence), puis, à l'intérieur de ces groupes, à apparier chaque ménage ayant besoin d'imputation (receveur) à un ménage dont le questionnaire est complet (donneur) et qui lui ressemble le plus par rapport à certaines caractéristiques (par exemple, revenu, nombre d'enfants, nombre d'adultes...). Les données provenant du donneur sont imputées au receveur si ces données permettent de satisfaire les règles de vérification de cohérence avec les données rapportées par le receveur.

Il est à noter que l'EDM recueille de l'information sur plusieurs aspects du budget d'un ménage. Or les questionnaires ne sont pas imputés en bloc, mais plutôt en sections correspondant en général aux sections du questionnaire, c'est-à-dire par groupe de questions ayant des relations entre elles. Ceci permet de maximiser le nombre de donneurs potentiels puisque si un ménage n'a qu'une seule question sans réponse, il peut servir de donneur pour les sections auxquelles il a complètement répondu. Cette approche entraîne la possibilité qu'un ménage soit imputé par plus d'un donneur. Ceci est minimisé par le fait

qu'on recherche le ménage le plus semblable possible selon certaines caractéristiques qui sont souvent les mêmes d'une section à l'autre. Il est important de noter que les questions d'une même section sont toutes imputées par le même donneur, ce qui permet de conserver les relations entre les questions.

5.3 L'identification des données utilisables

Les données de certains ménages pour lesquels le questionnaire est au moins partiellement complet peuvent être rejetées lors du traitement des données. Il existe deux causes principales de rejet. D'abord lorsqu'une partie importante des questions sur le revenu ou des questions sur les dépenses ont été laissées sans réponse, le questionnaire est classé incomplet et n'est pas utilisé. L'autre source de rejet correspond aux questionnaires qui ont été traités (vérification des règles de cohérence et imputation des données manquantes s'il y a lieu) mais pour lesquels la différence entre les rentrées d'argent (revenus et autres montants reçus par le ménage) et les sorties d'argent (dépenses plus la variation nette de l'actif et du passif) déclarés par le ménage est supérieure à 20%.

Après avoir été identifiées, les données utilisables sont pondérées pour produire par la suite les estimations.

6. LA PONDÉRATION ET L'ESTIMATION

Les estimations sont fondées sur le principe que chaque ménage de l'échantillon représente un certain nombre de ménages de la population cible, telle qu'elle a été définie dans la section 2. On attribue donc, à chaque ménage répondant, un poids d'enquête qui indique combien de ménages de la population sont représentés par ce ménage. Ce poids d'enquête est généralement le produit de trois facteurs : un poids de sondage, qui incorpore les données du plan de sondage; un facteur d'ajustement pour la non-réponse qui vise à compenser les ménages non-répondants et finalement un facteur d'ajustement pour rajuster l'échantillon en fonction de caractéristiques provenant de sources autres que l'enquête. On tient compte également, dans le calcul du poids d'enquête, d'un facteur d'ajustement pour les données influentes mais ce facteur affecte très peu de ménages.

6.1 Le poids de sondage

Le poids de sondage d'un ménage correspond à l'inverse de la probabilité qu'il soit inclus dans l'échantillon. Comme l'EDM est une enquête probabiliste, chaque ménage de la population cible a une probabilité connue d'être choisi dans l'échantillon. Si par exemple la probabilité de sélection d'un ménage est 1 sur 200, ce ménage aura un poids de 200.

Le plan de sondage, pour une répartition de l'échantillon donnée, permet de déterminer le poids de sondage. L'EDM utilise le plan de sondage de l'EPA qui est un plan de sondage autopondéré dans chaque strate c'est-à-dire que les poids de sondage établis lors de l'élaboration du plan sont égaux dans chacune

des strates. Dans la mesure où le plan de sondage et la répartition de l'échantillon ne sont pas modifiés, les poids initiaux pourraient être utilisés. Toutefois les étapes de stabilisation et de sous-échantillonnage décrites dans la section 3.3 modifient les probabilités de sélection initiales. Les poids de sondage de l'EPA sont donc ajustés pour tenir compte de ces modifications.

Comme l'échantillon de l'EDM correspond à un sous-ensemble des 6 groupes de renouvellement de l'EPA, on détermine les poids de sondage de l'enquête en ajustant les poids de sondage de l'EPA en fonction du nombre de groupes de renouvellement retenus. Ce dernier facteur peut varier d'une strate à l'autre puisque le nombre de groupes de renouvellement retenu dans chaque strate est déterminé de sorte à satisfaire les besoins spécifiques à l'EDM en terme de répartition de l'échantillon.

6.2 L'ajustement pour la non-réponse

Lorsque le répondant a omis de répondre à certaines questions seulement, les données manquantes sont imputées selon les méthodes décrites dans la section 5.2. Pour compenser la non-réponse totale, c'est-à-dire lorsqu'il a été impossible de contacter le ménage, lorsque celui-ci a refusé de répondre ou lorsque les données qu'il a fournies ne peuvent pas être utilisées, on ajuste les poids.

L'ajustement des poids pour la non-réponse est fondé sur le principe que les ménages répondants peuvent être utilisés pour représenter tous les ménages: les répondants et les non-répondants. Afin de procéder à ce rajustement, l'échantillon est d'abord subdivisé en classes de non-réponse définies de sorte à augmenter les chances que les répondants possèdent des caractéristiques semblables aux non-répondants.

Les classes de non-réponse correspondent à différents niveaux d'urbanisation dans chaque province ou territoire sauf au Québec, en Ontario et en Colombie Britannique, où les provinces sont d'abord subdivisées en deux ou trois régions infra-provinciales. Les niveaux d'urbanisation sont généralement les suivants: la principale région métropolitaine, les régions urbaines comprenant de 100 à 500 mille habitants, les autres régions urbaines plus petites et finalement les régions rurales ou éloignées. Dans certaines régions ou provinces, il peut être nécessaire de regrouper certains niveaux parce que l'échantillon est trop petit. Dans chaque territoire, il n'y a que deux classes de non-réponses: la ville principale et le reste de la population cible.

Les strates de ménage à revenu élevé constituent également des classes de non-réponse particulières dans chacune des provinces où elles existent. On notera que les secteurs de non-réponse ne se chevauchent pas et que, réunis, ils couvrent l'ensemble de la population cible.

Dans chaque classe de non-réponse, un facteur de compensation pour la non-réponse correspondant à l'inverse du taux de réponse pondéré de la classe est calculé. Ce facteur est, en d'autres mots, le rapport du nombre de ménages échantillonnés, pondérés par l'application du poids de sondage afin qu'ils représentent les ménages de la classe, au nombre de ménages répondants

pondérés. Pour éviter que les facteurs d'ajustement pour la non-réponse ne soient trop élevés, on regroupe certaines classes de non-réponse lorsque le facteur d'ajustement est supérieur à 2.

6.3 L'ajustement à l'information auxiliaire

En principe, en multipliant le poids de sondage par l'ajustement de non-réponse, on peut produire les estimations. Toutefois, il est possible d'utiliser des données auxiliaires sur la population cible pour améliorer les estimations de l'enquête. Si ces données auxiliaires sont corrélées avec les principales caractéristiques mesurées par l'enquête, des estimations plus fiables peuvent être produites. Par exemple, les dépenses des ménages sont corrélées avec la taille du ménage. Une mauvaise répartition de l'échantillon par rapport à la taille du ménage aurait un impact sur les estimations de dépenses. À partir de données auxiliaires sur le nombre de ménages selon la taille, on peut ajuster les poids de sorte à obtenir la véritable distribution du nombre de ménages.

Dans l'EDM, on utilise plusieurs sources de données auxiliaires pour ajuster les poids. D'abord, les estimations post-censitaires de population, produites par la division de démographie de Statistique Canada, fournissent des comptes du nombre de personnes selon différents groupes d'âge et selon le sexe par province ou territoire. Ces comptes correspondent à des projections de la population à une période donnée basées sur les données du recensement et sur de l'information provenant de dossiers administratifs tels que les naissances, décès, immigration, émigration, etc. Après avoir ajusté ces comptes de sorte à ce qu'ils reflètent la population cible de l'EDM à la fin de l'année de référence de l'enquête, les estimations pour dix-huit différents groupes d'âge et de sexe pour chacune des provinces[8] sont utilisées pour ajuster les poids. Les comptes démographiques du nombre de personnes de 18 ans ou plus et du nombre de personnes de moins de 18 dans certaines régions métropolitaines[9] sont également utilisés.

Les estimations du nombre de ménages selon la taille (une, deux ou trois personnes et plus) pour chaque province ou territoire ainsi que du nombre de ménages selon certains types de ménages sont également utilisées pour ajuster la représentativité de l'échantillon dans ces groupes. Pour les types de ménage, on utilise plus spécifiquement le nombre de ménages constitués d'une famille mono-parentale et le nombre de ménages composés de parents avec des enfants jamais mariés.

Dans le but de corriger certains problèmes observés au niveau de la distribution des revenus des répondants à l'enquête, les comptes du nombre d'individus dans certaines classes de revenu provenant de sources administratives servent également à ajuster les poids de l'EDM. Les données correspondent au nombre d'individus ayant perçu des revenus en salaires et traitements tel que rapporté par les employeurs. Six classes de revenus sont utilisées et les bornes de ces

[8] Dans les territoires, on utilise seulement quatre groupes : deux groupes d'âge croisés selon le sexe.

[9] Il s'agit des régions métropolitaines suivantes : St. John's, Halifax, Saint John, Québec, Montréal, Ottawa, Toronto, Winnipeg, Regina, Saskatoon, Calgary, Edmonton, Vancouver et Victoria.

classes sont basées sur les percentiles de la distribution suivants : 25, 50, 60, 75, 90 et 95[10]. Comme les données des fichiers administratifs de l'année de référence ne sont pas disponibles au moment où est effectuée la pondération de l'EDM, les comptes par classe sont projetés à partir des comptes des fichiers administratifs de l'année précédente et des tendances observées dans la distribution des individus selon le salaire dans l'EPA.

L'ajustement des poids pour tenir compte de tous les comptes décrits ci-dessus est effectué simultanément en utilisant une variante de l'estimateur de régression généralisé (ERG) fondée sur la méthode de pondération proposée par Lemaître et Dufour [4]. Cette méthode permet d'obtenir la concordance des estimations de l'enquête avec les estimations provenant des sources auxiliaires et assure en même temps, qu'après l'ajustement, tous les membres du ménage auront encore un poids identique. Le facteur d'ajustement calculé par l'ERG est ensuite appliqué au poids de sondage et au facteur d'ajustement pour la non-réponse pour produire le poids final du ménage.

6.4 L'ajustement pour les données influentes

Comme les dépenses suivent une distribution très asymétrique, les enquêtes sur les dépenses sont sujettes à la présence de valeurs extrêmes dans l'échantillon qui, combinées à un poids élevé, peuvent faire en sorte qu'un ménage a une contribution démesurée aux estimations. Les estimations de totaux et de moyennes, principalement les estimations provinciales ou toutes les estimations applicables à un sous-ensemble de la population, sont grandement affectées par la présence de telles données influentes.

Pour minimiser l'impact négatif de ces données influentes sur les comparaisons inter-provinciales et sur les estimations de tendance, il arrive qu'on ajuste le poids de certains ménages de sorte à réduire leur contribution aux estimations. Pour effectuer un tel ajustement, on utilise, comme information auxiliaire, la distribution des revenus des individus telle que fournie par les données fiscales des particuliers. Comme les dépenses totales sont très liées au revenu, la correction aura pour effet de diminuer l'impact sur les estimations des dépenses totales.

L'approche consiste à détecter les quelques individus qui ont une contribution importante (généralement de plus de 1%) aux estimations provinciales du revenu total. Si nécessaire, on ajuste par la suite le poids du ménage de ces individus, de sorte à ce que l'estimation du nombre d'individus avec un revenu de cette ampleur ne dépasse pas le nombre obtenu de la distribution des revenus des données fiscales des particuliers. Par la suite, l'ajustement à l'information auxiliaire, décrit précédemment, est appliqué à nouveau pour garantir la cohérence.

Il est important de noter que l'ajustement pour les données influentes, puisqu'il est axé sur les données très extrêmes, affecte le poids de très peu de ménages (généralement moins de 5 pour l'échantillon au complet).

[10] Pour certaines provinces, le 95^{ième} percentile est remplacé par le 98^{ième} ou le 99^{ième} percentile.

6.5 Les estimations

Dans l'EDM, les estimations des moyennes des dépenses par ménage sont produites en excluant les ménages qui ont existé seulement pendant une partie de l'année de référence, (tel que décrits dans la section 4.1). Ces ménages sont des ménages composés uniquement de personnes qui étaient membres d'autres ménages pendant une partie de l'année de l'enquête comme l'exemple des deux jeunes adultes qui vivent chez leurs parents et qui se marient et constituent un nouveau ménage pendant la période de référence. On y retrouve également les ménages constitués uniquement de personnes qui ont immigré au Canada pendant l'année de référence. Ces ménages ayant existé pour une partie de l'année correspondent à une très faible proportion de l'échantillon de ménages, soit moins de 4%.

Par contre, lorsque des estimations des dépenses totales de la population canadienne ou de sous-population sont produites, comme c'est le cas pour les besoins des comptes nationaux, tous les ménages de l'échantillon sont utilisés pour produire les estimations.

7. L'ESTIMATION DE L'ERREUR D'ÉCHANTILLONNAGE

Après avoir calculé les estimations, il faut en déterminer la fiabilité ou en d'autres mots estimer l'erreur d'échantillonnage associée à chaque estimation. L'erreur-type ou le coefficient de variation (qui est simplement l'erreur-type exprimée en tant que pourcentage de l'estimation) est la mesure courante de l'erreur d'échantillonnage. Elle correspond au degré de variation que l'on observe dans les estimations suite au choix d'un échantillon particulier plutôt qu'un autre. Comme l'EDM est une enquête probabiliste, on peut estimer l'erreur-type des estimations.

Dans l'EDM, on utilise l'estimateur jackknife pour estimer l'erreur-type. Cette méthode consiste à créer des répliques d'échantillon à partir des données de l'EDM. On crée autant de répliques qu'il y a d'unités d'échantillonnage du premier degré (UPD) en enlevant à tour de rôle une UPD de l'échantillon. Chaque UPD fait partie d'une strate et lorsque l'UPD est enlevée, les poids de sondage des autres UPD de la strate sont ajustés pour compenser ce retrait. On recalcule ensuite les estimations finales en appliquant les ajustements aux données auxiliaires décrits dans le section 6.3 aux répliques. En répétant cette opération pour chaque UPD de l'échantillon, on obtient autant d'estimations qu'il y a de UPD. La variabilité entre ces estimations sert à estimer l'erreur-type de l'estimation provenant de l'échantillon complet. La formule détaillée se trouve à l'annexe 1.

Il est important de noter que les estimations de l'erreur-type ou du coefficient de variation produites dans l'EDM ne tiennent pas compte du fait que certaines données ont été imputées : par conséquent, les CV calculés peuvent sous-estimer les valeurs réelles. Pour la plupart des variables de l'enquête, l'effet de l'imputation est minime. L'impact des données imputées pour chacune des

variables de dépenses est disponible dans le rapport de qualité de données spécifique à chacune des années d'enquête.

7.1 Modèle pour dériver une approximation du CV pour les estimations des domaines

Pour des raisons opérationnelles, il n'est pas possible de produire les CV pour toutes les caractéristiques collectées par l'enquête à tous les différents niveaux d'agrégations qui peuvent intéresser les utilisateurs, par exemple, par quintile de revenu, par type de ménage, par niveau d'urbanisation, par mode d'occupation du logement ou pour certaines régions métropolitaines. L'approche qui est suggérée aux utilisateurs de l'EDM est de calculer une approximation du CV en utilisant une relation entre le nombre de ménages de l'échantillon qui ont déclaré des dépenses pour une catégorie et le CV à un niveau agrégé (généralement au niveau national). Cette relation, basée sur la tendance du CV à croître proportionnellement à une diminution de la racine carrée du nombre de ménages déclarant une dépense, est illustrée dans l'annexe 2.

7.2 Modèle pour dériver une approximation du CV à partir du fichier de microdonnées

Les utilisateurs du fichier de microdonnées peuvent se servir d'une autre approche pour dériver une approximation du CV des estimations, qui sera généralement plus performante que celle décrite ci-dessus. Cette approche, relativement simple à appliquer, est décrite plus en détails dans la référence [5]. Elle est utilisable seulement à partir du fichier de microdonnées puisqu'il est nécessaire d'avoir les données et les poids de chaque ménage pour calculer cette approximation.

8. LA SUPPRESSION DE DONNÉES ET LA CONFIDENTIALITÉ

Certaines mesures sont prises pour assurer que les estimations produites à partir de l'EDM sont suffisamment fiables pour être publiées et que l'anonymat des ménages répondants est respecté.

8.1 La suppression des données non fiables dans les tableaux d'estimations

Comme le coefficient de variation est un indicateur de la fiabilité des données, idéalement on l'utiliserait pour déterminer si les estimations devraient être publiées ou non. Les estimations dont le CV est estimé à plus de 33% ne sont pas suffisamment fiables pour être publiées.

Par contre, le grand nombre d'estimations produites à partir de l'EDM fait en sorte qu'il n'est pas possible de calculer les CV pour chacune de ces estimations. Une étude effectuée à partir de données de l'enquête sur les

dépenses des familles a démontré que les CV atteignent en général environ 33% lorsque le nombre de ménages qui déclarent une dépense s'approche de 30. Cette règle est donc appliquée pour déterminer si les estimations de l'EDM peuvent être publiées ou non. Comme il s'agit d'une règle approximative, certaines estimations seront publiées même si le CV est supérieur à 33% et certaines estimations ne seront pas publiées malgré un CV inférieur à 33%. Certaines évaluations de la performance de cette règle se trouvent dans le rapport sur la qualité des données de 1997 [3].

On doit noter que même si les estimations de dépenses moyennes pour un certain type d'achat ne sont pas diffusées parce qu'ils ont été déclarés par moins de 30 ménages, les données sont retenues dans les estimations des composantes agrégées.

8.2 La confidentialité des fichiers de microdonnées

Quoiqu'un fichier de microdonnées à grande diffusion soit produit à partir des données collectées par l'EDM, celui-ci est différent de celui utilisé par Statistique Canada pour la diffusion des estimations. Ces différences proviennent principalement d'une série de mesures prises pour protéger l'anonymat des ménages qui ont répondus à l'enquête.

9. LES CHANGEMENTS DANS LA MÉTHODOLOGIE DE L'ENQUÊTE

La mise en place d'une enquête annuelle sur les dépenses des ménages a permis d'avoir des estimations sur les dépenses plus fréquentes et plus fiables, particulièrement au niveau des provinces puisque l'échantillon total a été augmenté et la répartition a été modifiée. Un nouveau questionnaire, beaucoup moins détaillé que celui de l'enquête sur les dépenses des familles a également été élaboré pour l'EDM de 1997. La méthodologie de l'enquête est toutefois demeurée assez similaire d'une année à l'autre à l'exception de la stratégie d'ajustement aux données auxiliaires dans la pondération.

Dans l'enquête de 1999, les projections démographiques basées sur le recensement de 1996 ont remplacé celles provenant du recensement de 1991 utilisées dans les enquêtes précédentes. De plus, la stratégie de pondération a été modifiée dans le cadre d'un projet qui vise à harmoniser les ajustements aux données auxiliaires dans les enquêtes de Statistique Canada liées à la statistique du revenu. Pour l'EDM, les principaux changements ont été l'utilisation de beaucoup plus de groupes (selon l'âge et le sexe) pour les comptes démographiques, l'ajout de comptes au niveau des types de ménages ainsi que pour certaines classes de revenu en salaires et traitements. Tous ces changements sont à la source de la révision historique des estimations de l'EDM 1997 et 1998 ainsi que de l'EDF 1996 et 1992 pour assurer la comparabilité dans les analyses des tendances.

BIBLIOGRAPHIE

[1] *Guide de l'utilisateur de l'enquête sur les dépenses des ménages*, Statistique Canada (disponible pour chaque année d'enquête), N° 62F0026MIF au catalogue

[2] *Méthodologie de l'enquête sur la population active du Canada*, N° 71-526-XPB au catalogue

[3] *Enquête sur les dépenses des ménages de 1997 – Indicateurs de qualité des données*, Division des méthodes d'enquêtes auprès des ménages (DMEM), Document interne, Statistique Canada (aussi disponible pour chaque année d'enquête), N° 62F0026MIF au catalogue

[4] Lemaître et Dufour (1987), *Une méthode intégrée de pondération des personnes et des familles, Technique d'enquête*, Vol.13, n° 2, pp.211-220, Statistique Canada

[5] Beaumont, J.-F. (2000), *Estimation de variance pour un fichier de microdonnées à grande diffusion provenant d'une enquête complexe*, Document de travail de la DMEM, HSMD-2000-002F/A, Statistique Canada

ANNEXE 1

Formule pour le calcul de la variance des estimations à partir du jackknife

Dans la procédure jackknife pour estimer la variance des estimations, on mesure la variabilité entre les estimations à l'aide de la formule suivante:

$$Var(\hat{Y}) = \sum_{h=1}^{H} \frac{n(h)-1}{n(h)} \sum_{i=1}^{n(h)} (\hat{Y} - \hat{Y}_{(hi)})^2$$

où

n(h) est le nombre d'UPE dans la strate h

$\hat{Y}_{(hi)}$ est l'estimation de Y quand l'UPE i de la strate h est enlevée.

L'erreur-type est la racine carré de la variance.

ANNEXE 2

Formule d'approximation du CV pour un domaine (un sous-groupe de la population)

Si CV(Y) représente le CV pour l'estimation de la moyenne par ménage d'une certaine caractéristique pour toute la population, alors on peut calculer une approximation du CV de l'estimation de cette caractéristique pour un domaine (que l'on peut considérer comme un sous-groupe de la population tel qu'un type de ménage, un quintile de revenu, un niveau d'urbanisation,...) à partir de l'équation suivante :

$$CV\ (Y_d) = CV\ (Y) \times \sqrt{\frac{nP}{n_d P_d}}$$

Où

n : *le nombre de ménage dans l'échantillon*
P : *l'estimation de la proportion des ménages déclarant une valeur > 0 pour cette caractéristique dans la population*
n_d : *le nombre de ménage de l'échantillon dans le domaine d*
P_d : *l'estimation de la proportion des ménages déclarant une valeur > 0 pour cette caractéristique dans le domaine d*

Généralement on utilise le CV, la taille n et la proportion P au niveau national pour calculer les approximations pour les différents domaines. Dans le cas où on cherche à calculer une approximation du CV pour une région métropolitaine, on peut utiliser ces valeurs au niveau provincial puisque le domaine est entièrement contenu dans une seule province et que les CV provinciaux seront publiés pour l'EDM.